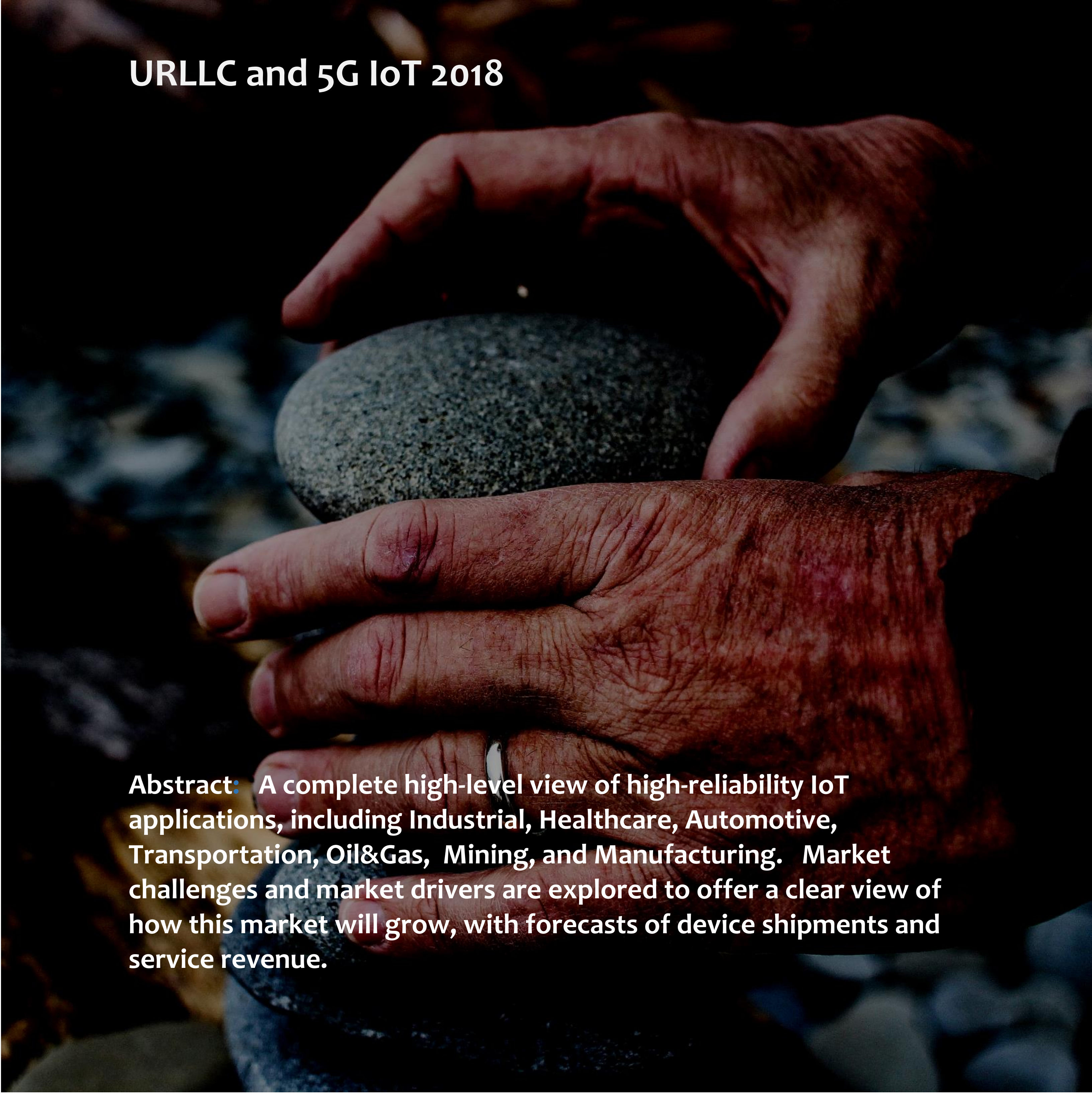
URLLC and 5G IoT 2018



Abstract: A complete high-level view of high-reliability IoT applications, including Industrial, Healthcare, Automotive, Transportation, Oil&Gas, Mining, and Manufacturing. Market challenges and market drivers are explored to offer a clear view of how this market will grow, with forecasts of device shipments and service revenue.

Joe Madden

November 2018

MOBILE EXPERTS

MOBILE EXPERTS

MEXP-URLLC-18

TABLE OF CONTENTS

CHARTS

FIGURES

1. EXECUTIVE SUMMARY

Early in the development of 5G, the industry defined three broad areas of market development: Ultra broadband, Critical IoT, and Massive IoT. All three of these areas have been treated as important possible growth markets, so the radio standard has been developed with enough flexibility to support all three cases.

RAN equipment vendors and chipset suppliers are investing a great deal of time and energy to develop low-latency, high reliability communications. The industry as a whole is investing an estimated $1B during 2018 alone, to make 5G capable of URLLC: Ultra-reliable, low latency communications. But is this investment justified?

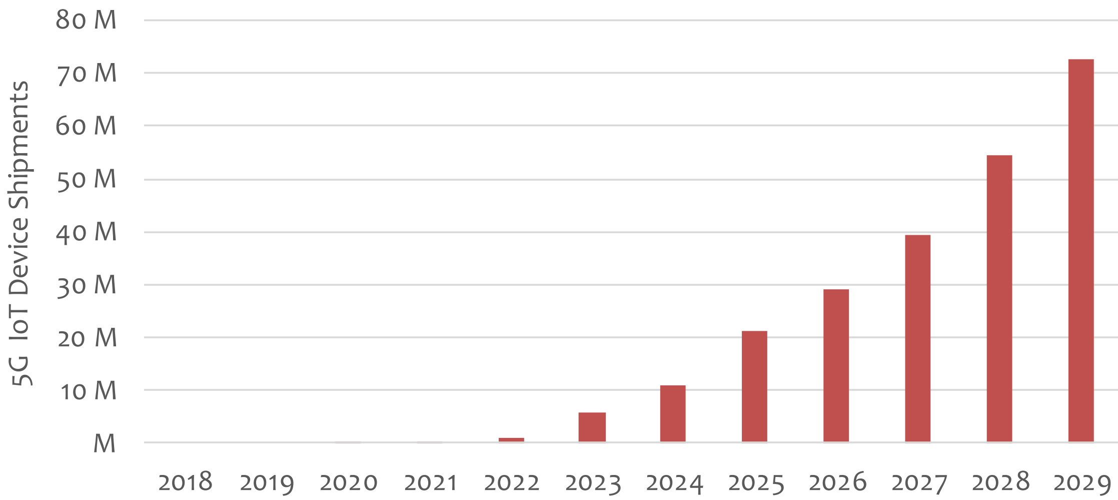
Our view is that, for the next five years, the market is far smaller than most people expect. Many applications are considered, but in some cases the customers are not ready to pay premium prices for lower latency or guaranteed coverage.

Examples of applications that are not realistic include:

- Self-driving cars: Our interviews with GM, Ford, Honda, Toyota, and several others are clear: 10-20 ms latency will work for them, because any time-critical decisions are made onboard, and V2V or V2X communications will be used for long-range information, not for braking the car.
- Drones: So far, drones are also designed so that critical control is achieved on board. Delivery of packages and other drone applications are automated so that a low-latency link to a human pilot is not considered necessary for mass deployment.
- Remote Surgery: This popular idea has problems with market size…a few units may be sold but only for specialty niche cases.

On the other hand, a few examples show promise for premium IoT pricing:

- Industrial operations such as control of the electrical grid may require wide-area coverage, low latency, and guaranteed high reliability. The market is unclear so far on whether this would be implemented through a mobile operator, or as a Private 5G network.
- Manufacturing robot controls can be a promising area. Wireless control in the factory has never taken off due to poor reliability using Wi-Fi and other unlicensed formats…but factories can realize a positive ROI from the added flexibility of 5G controls and data collection.
- Virtual and augmented reality applications can require low latency, so as an improved human/machine interface there is long-term potential here for unknown use cases. In the URLLC context, some combination of augmented reality and robotic control may become important in either consumer or industrial markets….but the specific applications are not clear today.



Source: Mobile Experts

Chart 1: Global 5G IoT device shipments, 2018-2029

Overall, we do not expect this market to rise rapidly. In the next five years, 5G URLLC device shipments will grow to only about 4 million per year, and service revenue will be a trivial percentage of total revenue for mobile operators. Private 5G networks are likely to capture a significant share of the market, as industrial operations prefer to fully control their data and the networks that touch the data.

Fundamentally, our conclusion is that the billion-dollar investment that the mobile industry is making into 5G URLLC is not going to pay off right away. There may be a payoff in long term applications, but we cannot make firm predictions about them now.

2. MARKET OVERVIEW

As the number of smartphone subscribers peaks and revenue growth stagnates in the mobile industry, the operators and OEMs are all looking for the "next big thing". In particular, the OEMs have invested heavily in developing 5G with flexibility for high reliability and low latency.

The question is…now that 5G URLLC exists, will anybody pay for it? Mobile Experts examined two primary business models. Note that both of these business models rely on enterprises to pay the bills at the end of the day, but the approach is quite different in each case.

1. Mobile operators can deploy large-scale, wide area networks with 5G wireless services in order to pursue "nationwide" opportunities for IoT connectivity. Consumers have little need for guaranteed reliability or super-quick latency, so in this scenario the operators are planning to pursue large enterprises with premium IoT connectivity. In some cases the operator anticipates deployment of small cells inside a building or covering an industrial operations area, so investment may be highly targeted for specific customers.
2. Enterprises can invest directly in their own infrastructure. In this case, the enterprise will need spectrum that is designated for shared or unlicensed use…or will need to find a way to license spectrum for themselves.

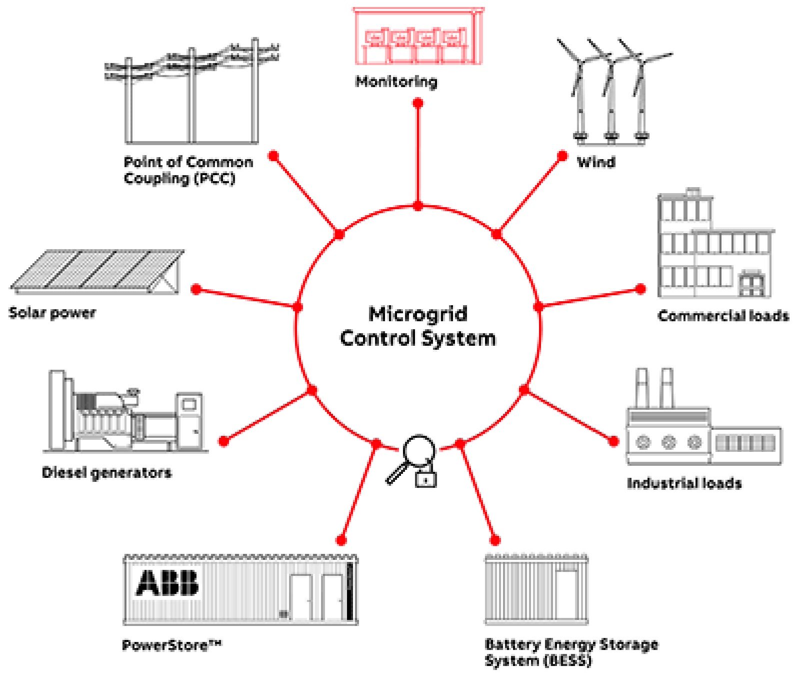
Mobile operators have been unable to serve this market with 2G, 3G, and 4G technologies so far. Over the past twenty years, operators have discussed wireless support of industrial operations many times, but the discussion often stops when the industrial customer asks for guaranteed reliability and latency from the network. Electric utilities want to monitor transformers on the grid….but when the mobile telco refuses to guarantee 99,999% availability, the discussion ends.

Mobile Operators and Wide Area Networks

Every new mobile service starts with wide-area coverage from macro base stations. We expect URLLC services to be available on licensed bands below 6 GHz, with the radio coverage shared with broadband services in those bands. Network slicing in the overall network would enable the URLLC service to take priority over the dynamic ups and downs of broadband users.

A few specific applications for wide-area URLLC have been widely promoted for the past four years:

- Mobile industry experts speculate that driverless cars will absolutely need < 1 ms latency and 99.999% reliability. The argument is that safety in the automotive market will justify wide-area network investment. However, in our interviews with GM, Ford, Honda, Toyota, Daimler, and Volkswagen, none of the automotive experts are convinced that URLLC is truly necessary and worth a billion-dollar investment. Their response, in short, is that important decisions will be made by the car using on-board sensors…and that RF connectivity will enhance information in long-range awareness of road conditions. In this case, 10-20 ms latency is fine.
- Robotic surgery is another wide-area application that has been very popular in presentations. But we have not been able to find any customers for this kind of product. The people with enough money to buy a robotic-surgery machine would absolutely prefer to visit a hospital, instead of experiencing a surgical procedure in the field. People that cannot get to a hospital are generally too poor to buy this machine. There will be exceptions in cases such as military medical units, so we project a small market that does not justify wide-area network investment.
- Drone control is one possible area, but increasingly these drones appear to be fully automated, so low-latency radio links are being ruled out as a requirement. We're watching this area to determine whether FAA or other regulation will require a drone network to enable positive control by a human pilot under key conditions….such a government mandate would change the requirements, but until that happens, drones may not drive significant network investment.
- Controlling the electrical grid is an application that could require very precise synchronization over long distances. Because networks don't support this use case today, we don't have a proof of concept network to validate it. However, in discussions with utilities and suppliers this area holds potential for both sensors and actuators that can change grid settings (throw switches) and react very quickly to changes. In particular, high-voltage infrastructure must be handled with care, and switches can only be thrown at "zero crossings" which can require timing control in the microsecond to millisecond range. One electric utility in San Diego is implementing LTE to shut off power "before the broken cable hits the ground", and has stated an intent to migrate to 5G for a higher safety margin to avoid starting wildfires.



Source: ABB

Figure 1. A sample diagram showing complexity of microgrid control.

Altogether, for wide-area URLLC services we expect only very limited success for the next five years. One reason is on the technical side: while the radio standards are now available, for a wide area service the radio latency comprises a small portion of overall latency. The radio itself may allow for a round-trip connection time of 1 ms, but 100 ms or more can be required to access a large-scale cloud service. This problem is surmountable, with heavy investment in Edge Computing platforms in local sites. However, our analysis illustrates the fundamental problem: Edge Computing at a tower-by-tower level in a nationwide network can be extremely expensive. Backing off to Edge Computing using regional data centers can be a good compromise, but of course as the computing platform is moved farther away from the client device, the latency increases and reliability becomes more difficult to guarantee.

In the long term, we expect that nationwide services for electrical grids, industrial machine automation, and other new ideas will be possible and profitable. But the level of investment involved will not happen quickly, as these new applications will take time to reach scale.

Even using some generous assumptions (for example, assuming that a nationwide URLLC service can address 100,000 devices immediately, at $100/month), we find that the payback time on a nationwide network investment can be more than five years. For this reason, we don't believe that operators will take a more passive approach to the investment: If 5GURLLC comes along with the 5G network for free, then they will try it.   But we should not expect an extra investment for this business case.

Latency Sensitive Application---Robotics				
Scenario	CAPEX for EC	Backhaul CAPEX Savings	Revenue Stream per month	Payback (months)
At Radio site	$8 M	$0 M	$5 M	3

Source: Mobile Experts

Figure 2.   ROI for Nationwide URLLC Services

Mobile Operators Investing in Localized Services

Leading mobile operators in countries such as the United States, Germany, UK, and Japan are now planning a more targeted approach at URLLC services.   Instead of relying exclusively on wide-area networks, these operators plan to target large enterprises with services for specific operational facilities.   In the USA, Verizon and AT&T are targeting Fortune 500 companies with industrial operations that stand to benefit from URLLC capability.

In this scenario, the investment decision can be very clear and specific. The operator would not necessarily deploy a nationwide network for billions of dollars to generate its first dollar of revenue. Instead, smaller investment increments can be considered that are tied directly to a specific customer.   Examples could include:

- Manufacturing facilities can replace Ethernet and Fieldbus cabling with wireless control of robotics.   This approach can allow the factory to scale up, reconfigure, or change operations quickly, and without disrupting the product flow.   The robotics case requires very quick response time in order to achieve high efficiency in the factory, as well as safety in an environment where robots and people work together.
- Oil refineries or other major outdoor industrial operations require real-time control of automated equipment.   The latency requirement is unknown today, but some scenarios have been discussed at a white-board level with energy suppliers, showing some interest. We don't see a lot of investment by energy companies to move in this direction so far.
- Mining operations are famous examples of Private LTE, where automated trucks and excavators move tons of ore around in a radio-controlled ballet.   Multiple Private LTE networks have been set up in Australia, Argentina, Russia, and other countries.   So far, most of these have relied on the remote nature of the mining operation to re-use mobile spectrum that otherwise would be underutilized in that area.

- Port operations use huge automated cranes to move large steel containers from ships to the docks, as well as automated trucks to move the freight around on the docks themselves. Each huge crane can have as many as eight cameras, streaming HD video… but also need to be controlled for safety as longshoremen need to work in the same areas as the cranes and trucks. Onboard computers can make many decisions, but in many cases the cranes and trucks must be controlled by a technician in the port offices. Human judgment cannot always be replaced by automation, so low latency is required for the control aspects of the crane or truck communications.

Source: Robin Summer

Figure 3. Industrial operations improve results with real-time control

For localized operations, the mobile operator might be able to add value, but we think it's more likely for the private enterprises to invest in their own equipment. As with the remote mining case, we find that large industrial firms are very clear about wanting control over their network, and want to keep their data safely within their own organization.

Private Enterprises Investing On-Premises

Because most industrial giants have serious and justified concerns about security, they invest heavily in IT infrastructure and physical assets to protect themselves. But a simple and less tangible way to improve security boils down to a simple principle: Keep your data contained, and fewer people will be able to access it.

So, despite a strong desire by the mobile operators to participate in IIoT, they will be shut out of many localized opportunities. The single leverage point that the operators have: They own the clean spectrum that is required to run effective high-reliability wireless services.

The spectrum challenge is real for enterprises. Several examples exist of railroads, electric utilities, and other industries that have acquired small blocks of spectrum. For example, the Tennessee Valley Authority (TVA) monitors their electrical grid using a 2 MHz wide spectrum at 220-222 MHz. Other narrow slices of spectrum are available in the 900 MHz range as well, including many train control systems. These bands are fairly narrow, making them difficult to keep "clean" from an interference point of view. The narrow band also makes these small blocks of spectrum less useful for video, voice, or other applications that require richer content.

In the end, a few enterprises have been able to license wider blocks from operators. Rio Tinto is one example, where in Australia they were able to license LTE spectrum that was not utilized in a remote part of Western Australia. But that doesn't work for a factory in an urban setting, so spectrum is a serious hurdle for enterprises to consider investing in this direction.

Source: Rio TInto

Figure 4. Rio Tinto's automated excavation operations with Private LTE

The CBRS frequency band is a game-changer in this regard. When the FCC finally completes an auction for Priority Access Licenses in the early 2020 timeframe, we expect them to reserve at least 30-40 MHz for local licenses. This means that the FCC has received a lot of requests from local industrial firms that want to bid for spectrum within their county.

3. TECHNOLOGY OVERVIEW

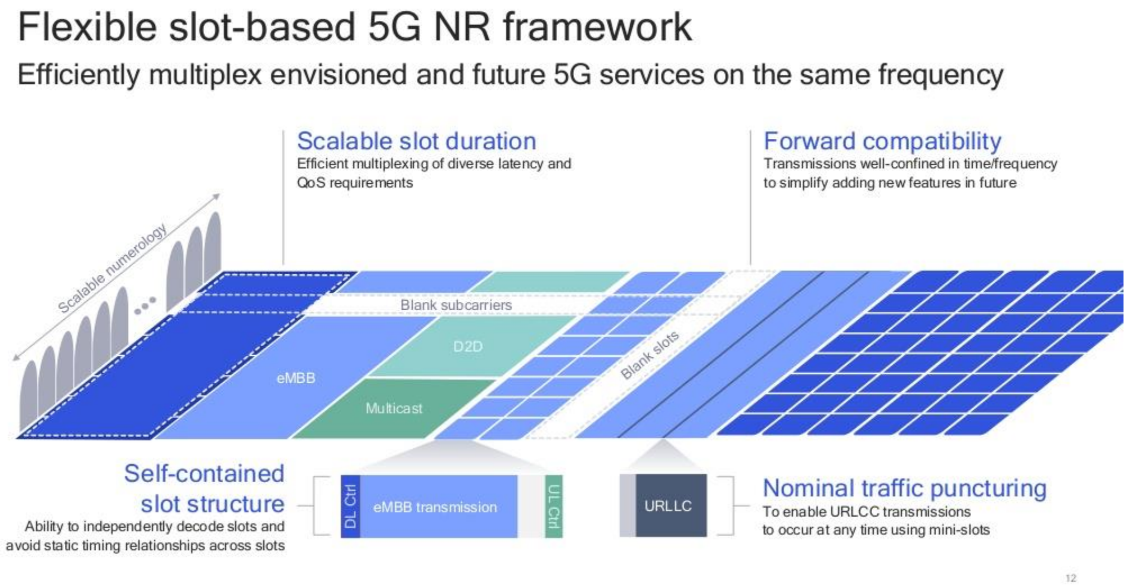
Ultra-Reliable Low Latency Communications (URLLC) is not necessarily the same thing as 5G Communications. We have low-latency control of systems in manufacturing, electric utilities, and many other areas today, using wired technology. In the manufacturing environment, Fieldbus systems and Ethernet variations have been adapted for very high reliability, and the latency of these wired communications protocols is very low, in the range of microseconds.

In addition, the building blocks of URLLC have been available in Release 14 of 3GPP for LTE, and network operators have had the tools to put together a version of a low-latency network with dedicated resources for a few years already. So what is different now? With Releases 15 and 16 of 3GPP standards, we anticipate that the cost of setting up a highly reliable network and achieving low latency will be lower.

1. The 5G NR wireless format includes inherent flexibility that allows for rapid two-way communications. The basic frame structure allows for a quick bi-directional exchange of control data to establish the link and the 5G data symbols follow immediately, so short messages can be completed very quickly.
2. Network Slicing allows for software-based priorities in the network, resulting in a capability to guarantee availability for the IoT service, no matter how much broadband traffic comes through.
3. Edge Computing is coming to the market at the same time as 5G technology. By collapsing the core network into more local data centers, and moving the cloud computing center closer to the user, the service can be quicker and more deterministic (more reliable).

5G NR

From the beginning, the RAN committees that developed 5G NR have been concerned with creating a frame structure that was flexible enough to handle broadband data and highly reliable, low-latency bursts.

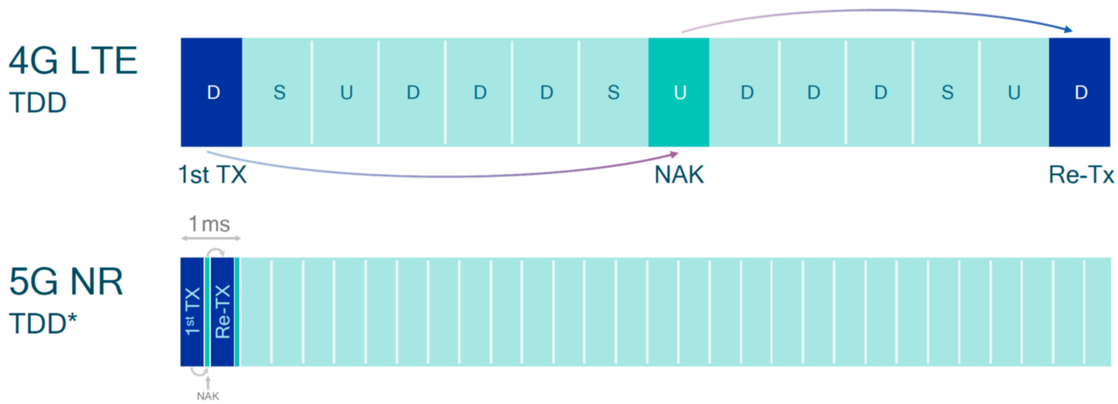


Source: Qualcomm

Figure 5. 5G NR frame structure and areas of flexibility

The 5G NR waveform has multiple improvements over LTE in terms of latency:

- The Transmission Time Interval (TTI) can be set to a much shorter time. Where LTE has a fixed TTI of 1 ms, in 5G the 'mini-slots' can be set to allow TTI in the range of 140 microseconds.
- The shorter slot lengths go along with higher 'numerology', which refers to the sub-channel spacing and basic setup of the OFDM access. The sub-channel spacing is typically locked into the hardware design, so changing the standard to allow for narrower sub-carrier spacings means that IoT can be quick while operating in a narrow band.
- The processing in 5G NR is quick, using only 1-2 symbols in the UE, and 7-14 symbols in the gNodeB. In this way, a message can be answered quickly.
- Grant-free uplink messages are possible with 5G NR, simplifying the control signals necessary to set up a link.
- Rapid Hybrid Automatic Repeat Requests (HARQ) are possible, greatly improving latency under marginal radio conditions.
- Puncturing is now available with 5G NR, so that other data traffic can be interrupted with mission-critical traffic without loss of continuity.



Source: Qualcomm

Figure 6. Low-latency HARQ will speed up 5G URLLC

Network Slicing

One problem that mobile operators have in selling services in the enterprise market is that operators cannot really guarantee reliability and latency today. When an IIoT customer asks the mobile operator to sign a Service Level Agreement (SLA) with performance guarantees, they don't have good tools to comply.

As 5G develops a bit farther in Release 16, we expect Network Slicing to be used as a way to solve this problem. In this approach, a customer can pay to take a priority position in the network, guaranteeing that critical traffic will be transmitted despite a heavy load from other users.

There are two potential issues with this feature:

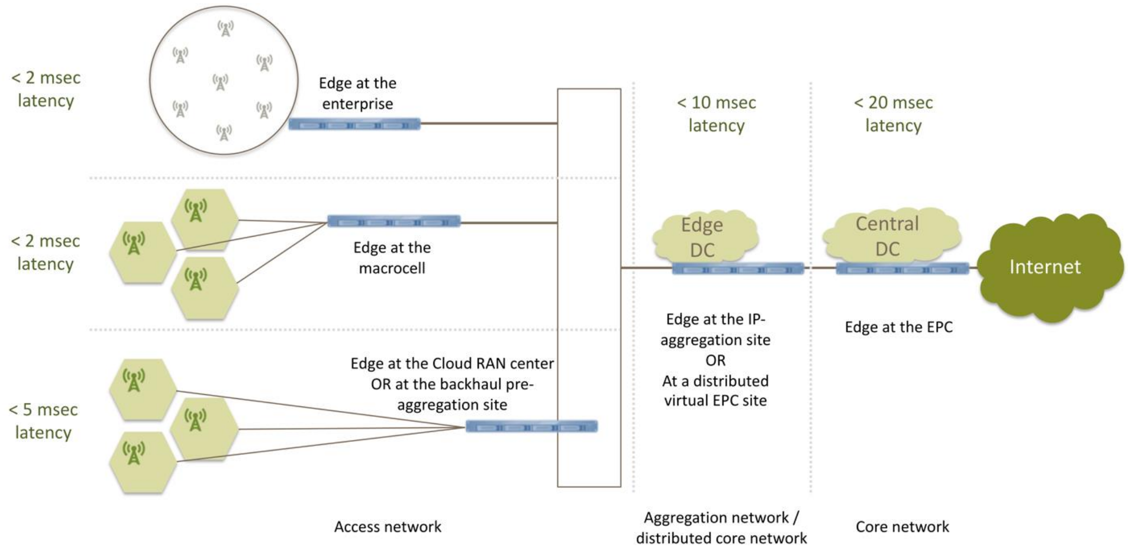
- Net Neutrality could be a problem here. So far, Net Neutrality has not been a major impact on the mobile industry, but "paying for priority" is prohibited under Net Neutrality regulations. Will these regulations really be enforced? We cannot adequately forecast the outcome here. It's true that broadband users are very unlikely to be harmed by little bursts of IoT traffic, so in this particular case we hope that long-term regulation will allow URLLC slices.
- Mobile operators are not keen to set up slices. Despite the potential revenue of URLLC services, video services, automotive services, and other possible business scenarios, the operations team at each major operator has expressed major reluctance. The reason is that planning, provisioning, and testing each network slice is a major manual exercise. A single network slice today would require man-years (!) to set up. Automation should dramatically reduce that, but automation will take time.

Source: Ericsson

Figure 7. Network slicing diagram

Edge Computing

Edge Computing, Mobile Edge Computing, Multi-Access Edge Computing, Fog Computing…..the market can be confusing regarding what is really happening. At a high level, we see the trend toward more distributed computing, with value in localized content and faster latency for some applications. All of the different names reflect nuances for various market segments, as each vertical market has a different definition of the "edge", and different economic calculations for where the computing should live.



Source: Mobile Experts

Figure 8. Edge Computing Alternatives

Our recent survey of mobile operators shows that they're investing in Edge Computing, setting up cloud computing centers of their own. AT&T is deploying 200 regional data centers, co-locating the core network and the cloud servers in position to offer about 10-20 ms latency for most applications. Other operators are looking at similar investments.

At the same time, most IoT developers prefer to create products based on Azure or AWS Greengrass, which means that the mobile operator edge clouds would not be in position to succeed. For this reason, we believe that the operator investments are likely to shift so that the operators begin to host AWS, Google, Azure, and others worldwide in order to achieve the benefits of hyperscale and the localization of close proximity to the core network and the end user.

In last year's analysis, Mobile Experts calculated the costs and benefits of placing Edge Computing resources in the radio towers, compared with regional data centers and more centralized schemes. The cheapest implementation can be a centralized core network and centralized cloud computing. But with higher latency and high transport costs, this is not ideal. The lowest latency and best performance comes with deployment in the radio, but the cost is staggering. AT&T, China Mobile, Verizon, and other operators are settling in between for their initial deployment.

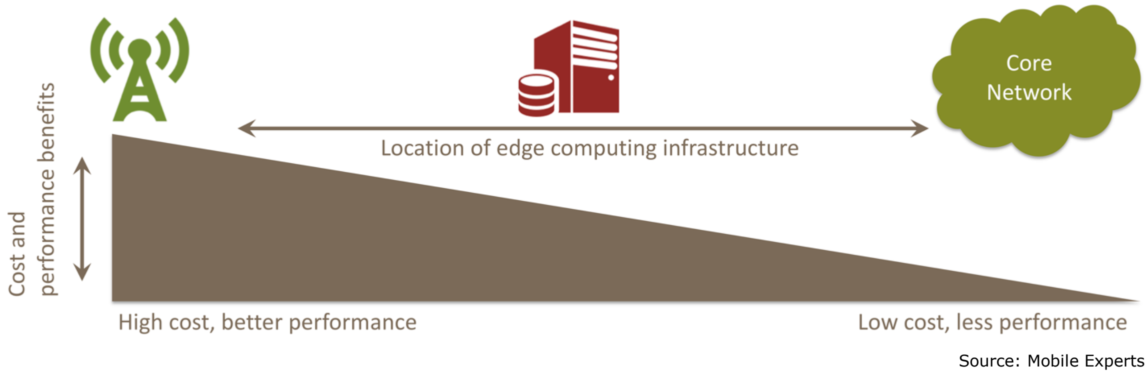


Figure 9. Cost and performance tradeoffs in Edge Computing

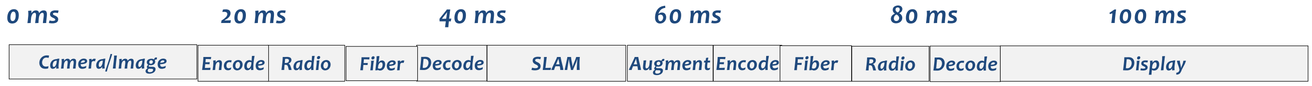
For URLLC services in an industrial setting, with Private 5G networks deployed on premises, we expect to see system integrators pulling together the players, including 5G network infrastructure that is blessed by an operator, as well as edge servers that are blessed by the selected Cloud service: AWS, Google, Azure, etc. Instead of a single company creating a proprietary product for the IIoT customer, a coordinated product will likely be required.

Latency

Let's be careful when we talk about latency. Many of the 5G demos illustrate "the need for 5G" by showing hundreds of milliseconds of latency in VR applications or IoT applications. In fact, the radio latency is a small percentage of the total.

Looking at one example of an augmented reality application, we can see that the majority of the latency is NOT related to the radio at all.

- About 18 ms is used by the camera, image processing, and storing the image in RAM on the mobile device;
- In this example, we assume a fast 5G radio with 1 ms latency… but each 1 MB image will require 8 ms to transfer (1 , for a statistical average of 9 ms.
- We assume about 100 mile distance between the local base station and a regional Edge Computing data center, equivalent to about 5 ms latency.
- Simultaneous Localization and Mapping (SLAM) requires about 15 ms for its computations, determining the position and orientation of the image.
- Augmentation/computing to enhance the image is assumed for this example to happen quickly, in about 8 ms.
- Then, the data makes a return trip through encoding/fiber/radio/decoding again;
- Finally, the RAM in the mobile device must load the image, and the image must be rendered, with about 25 ms for the image to refresh on the screen.
- Total "Glass to glass" time is more than 110 ms.



Source: TU Munich, Mobile Experts

Figure 10. Latency budget in an Augmented Reality example

Overall, for 110 ms of latency in this application, the radio latency only accounts for 8 ms in each direction. Clearly, 5G is not the limiting factor here.

In an application for industrial robotics, the latency budget stacks up a bit better. Assuming that the robotic devices are optimized for low latency in the cameras, memory, and on-board computing platforms, we should see a total control loop timing of less than 5 ms. The radio latency in this case can be fully completed within 1 ms, because we're dealing with small bits of data and short distances. We assume that the computing platform will be on-premises, connected via Ethernet to the 5G access points throughout the factory.

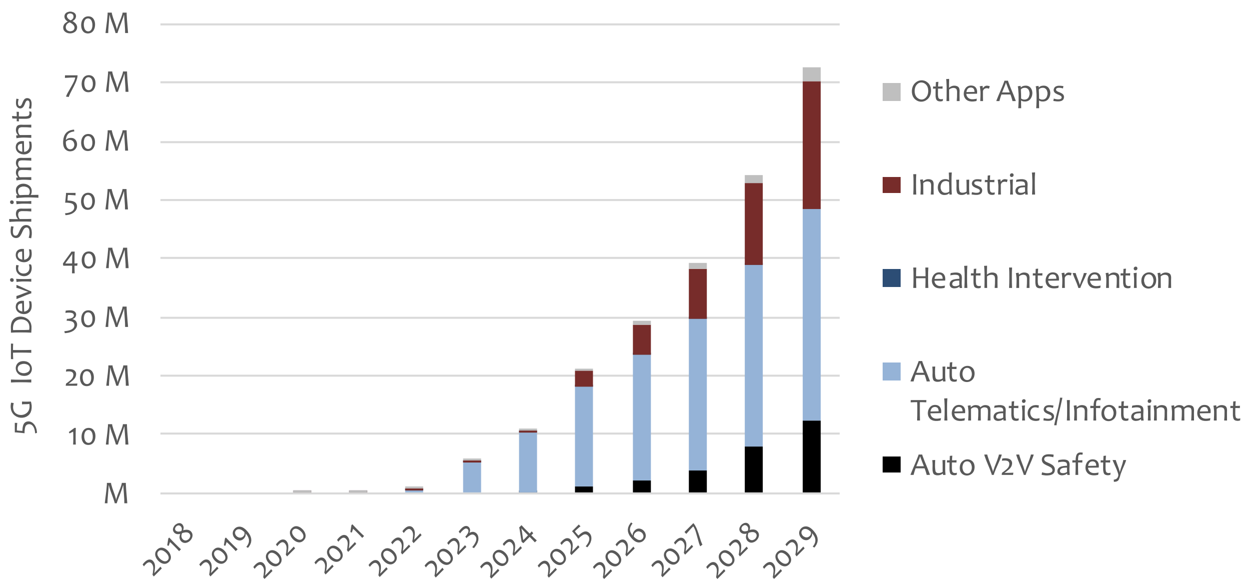
Note that the robotic arm itself may not react within 1-2 ms. Large robotic machines have mechanical factors which will limit the actual time to stop/start/turn in the range of 20-50 ms. Our example only reflects the 5ms round-trip time to sense a problem, and send a control signal to remedy the problem.

Source: Mobile Experts

Figure 11. Latency budget in a Robotic Control example

4. MARKET OUTLOOK

For this market, Mobile Experts has chosen to extend our normal 5-year forecast window to 10 years…because frankly the five-year forecast was too short to show the impact of new applications in this market. New industrial applications will take a long time due to long design cycles (5-10 years) and slow adoption pace in the industrial sector.

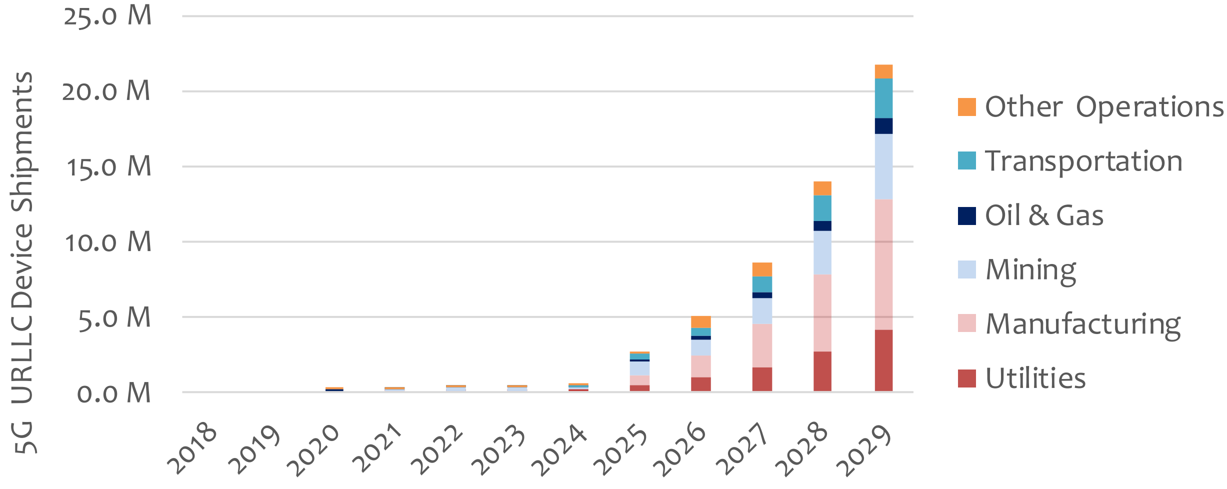


Source: Mobile Experts

Chart 2: 5G IoT Device Shipments, by vertical market, 2018-2029

For example, automotive design cycles last about 7 years, and then the car is expected to be used for another 20 years. Therefore we expect to see 5G growth in telematics, industrial, and other applications exceed 10 million units per year in the 2026-2029 timeframe.

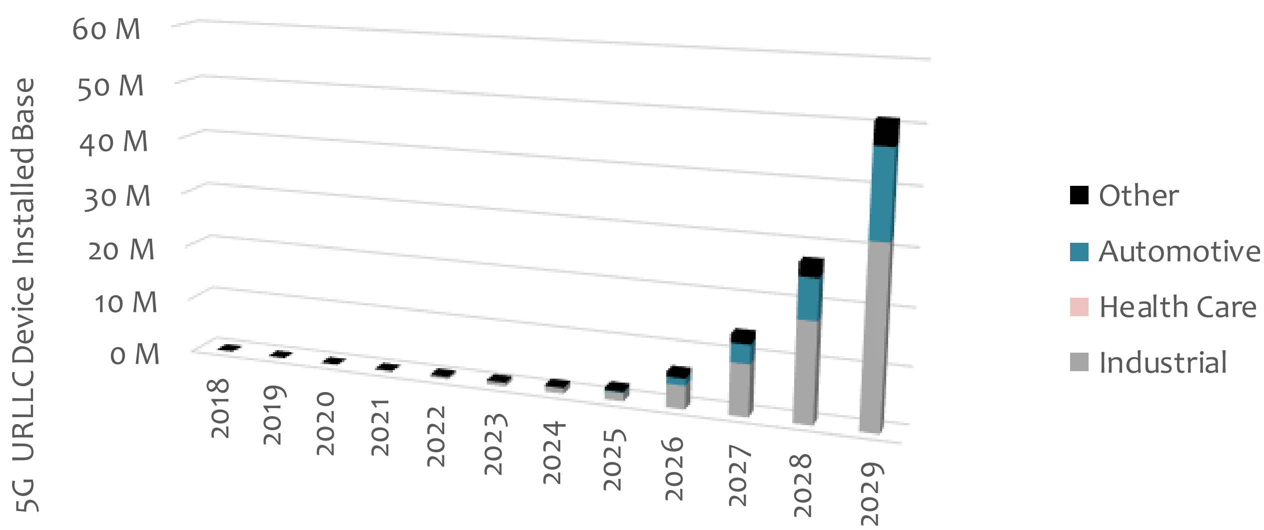
Out of these 5G applications, we should note that some of them (such as automotive telematics/infotainment) are not really URLLC applications, because they're intended more as a broadband modem and not for instant machine-type communications. As a result, our 5G URLLC forecast is smaller than the total number of 5G IoT devices.



Source: Mobile Experts

Chart 3: 5G URLLC Device Shipments, by vertical market, 2018-2029

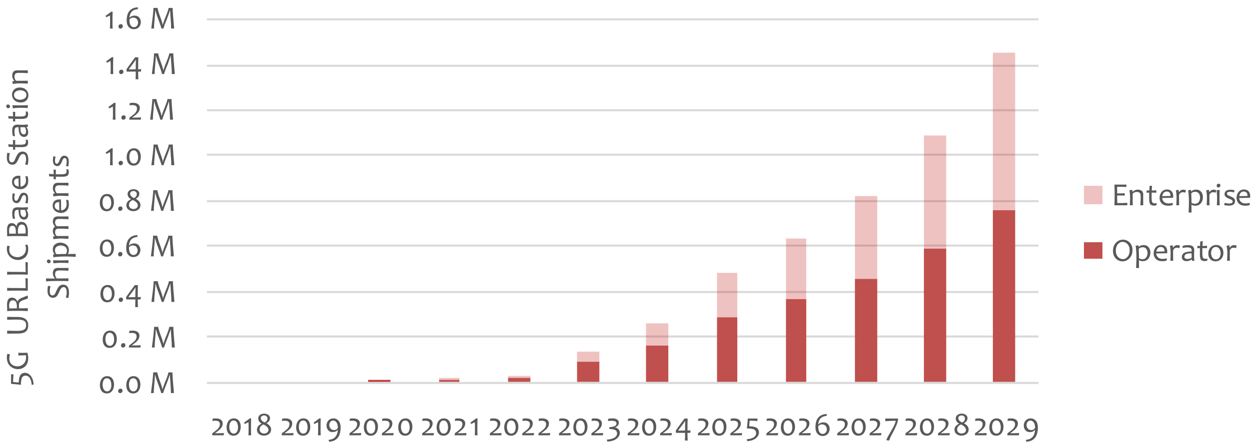
Service revenue related to URLLC devices will, of course, be limited by the number of devices in the market. The installed base of 5G URLLC devices will grow steadily, to over 30 million devices in 2029.



Source: Mobile Experts

Chart 4: 5G URLLC Device Installed Base, by application, 2018-2029

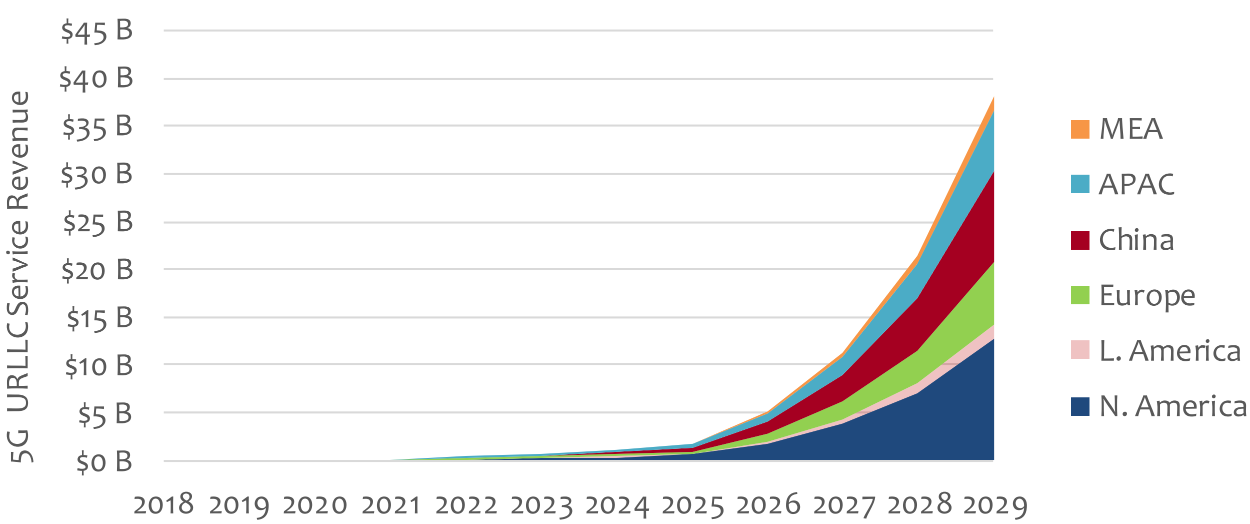
5G URLLC base stations will be deployed in two different ways: On a wide scale, a mobile operator will make a software update to the mobile broadband network, to set aside frequency/time slot resources for high-priority URLLC traffic. At a smaller location with limited need for mobility, enterprises would prefer to deploy their own networks, with spectrum that is somehow licensed over a small area or sub-leased from a mobile operator.



Source: Mobile Experts

Chart 5: 5G URLLC Base Station Deployment, Enterprise vs Operator, 2018-2029

The revenue per device should be high, with enterprises today expressing a willingness to pay hundreds of dollars per month. Assuming that pricing does not drop dramatically, we should have a service revenue market in the neighborhood of $5-10B by 2029. The profile of revenue is weak for the first 5-7 years of our forecast, because we expect the industrial applications to need time to incubate. After about 2026, adoption can grow quickly.



Source: Mobile Experts

Chart 6: 5G IoT Mobile Operator Service Revenue, 2018-2029

The profile of service revenue worldwide is fairly evenly spread according to our current view, because industrial and electrical utility applications are relevant on every continent.   Over time, the regional profile could change dramatically depending on what applications are most successful.

5. GLOSSARY

2G:	Second Generation Cellular
3G:	Third Generation Cellular
4G:	Fourth Generation Cellular
5G:	Fifth Generation Cellular
BBU:	Baseband Unit
CBRS:	Citizens Broadband Radio Service, a shared wireless broadband use of the 3550-3700 MHz (3.5GHz) band in the US
CPE:	Customer Premise Equipment (e.g., cable modem, broadband gateway)
CPRI:	Common Public Radio Interface
CRAN:	Centralized RAN
DWDM:	Dense Wavelength Division Multiplexing
EVM:	Error Vector Magnitude (metric highlighting modulation accuracy; lower % or higher decibel value implies better modulation)
GHz:	Gigahertz
HARQ:	Hybrid Automatic Repeat Request
I/Q:	In-phase and quadrature baseband signal
IoT	Internet of Things
IP:	Internet Protocol (or Intellectual Property)
LAA:	License Assisted Access
LTE:	Long Term Evolution
LWA:	LTE Wi-Fi Aggregation
MHz:	Megahertz
MIMO:	Multiple input, multiple output spatial multiplexing
ms:	Millisecond
OFDM:	Orthogonal Frequency Division Multiplexing
RAN:	Radio Access Network
RF:	Radio Frequency

ROI:	Return on Investment
RRH:	Remote Radio Head
SLA:	Service Level Agreement
TTI:	Total Time Interval
URLLC:	Ultra Reliable Low Latency Comms (< 1ms, 99.99% or higher rel)
V2V:	Vehicle to Vehicle
V2X:	Vehicle-to-Anything
W:	Watts of power
Wi-Fi:	Wireless Fidelity (unlicensed wireless communications)

6. METHODOLOGY

To create estimates and forecasts for URLLC networks and client devices, Mobile Experts relied on direct input from more than 20 industry sources, including input from multiple mobile operators and enterprises in key industry areas. The technology suppliers were very helpful in illustrating the technical details, but we relied exclusively on our interviews with industrial firms to build our assumptions concerning market adoption.

To define this market, Mobile Experts segregated the portion of the market where 1 ms radio latency would be required, or where at least 99.99% availability of the radio must be guaranteed. It's important to note the distinction between guaranteed availability and the "best effort" services available today. In most cases, the reliability requirement means that only locally deployed small cell networks will be capable of achieving the guarantee.